




Systematic Review of The Challenges in Implementing Data Segmentation with Machine Learning

Benjamin J. Garcia¹ ; Jorge L. Ruiz² ; Luis C. Rada³ 

^{1,2,3}Universidad Tecnológica del Perú, Peru, u20219970@utp.edu.pe, c02047@utp.edu.pe, c18380@utp.edu.pe

Abstract— *This study identifies the main challenges in implementing data segmentation using machine learning techniques. A systematic literature review was carried out using the PICO methodology and the PRISMA framework, which allowed the selection of 44 relevant articles. The predominant methods include Deep Learning techniques, Ensemble Learning and traditional classification approaches, applied in domains such as telecommunications, health and cybersecurity. Among the highlighted challenges are the high complexity of the data, the presence of noise, the inconsistent quality of the records and the difficulty in integrating heterogeneous sources. Despite the progress made, limitations persist in terms of scalability and the absence of standardized methodological frameworks. This study is very useful for researchers oriented to improve the precision and efficiency in data segmentation in environments with large volumes of information. The development of adaptive methodologies and the establishment of standards that facilitate the transfer of knowledge between sectors are proposed.*

Keywords—Data segmentation, deep learning, ensemble learning, implementation challenges, machine learning.

I. INTRODUCTION

With the growth in data complexity and volume, segmentation using machine learning techniques has become increasingly important, facing critical challenges such as data quality, the presence of noise, and the need to integrate robust models that adapt to dynamic environments. Recent research has addressed these difficulties in various contexts. For example, [1], [2] highlight the importance of adapting models to local data properties, such as noise distribution, which affects segmentation accuracy.

Likewise, [3], [4] emphasize that more robust methods, such as improvements to the K-Means algorithm, can reduce prediction errors, although they sometimes compromise clustering quality. Additionally, [5] demonstrates how temporal segmentation optimizes energy predictions, while [1] uses it to detect changes in machine lifecycles. However, the presence of noise makes it difficult to accurately identify phase boundaries in both cases.

Despite these advances, significant gaps in the literature persist, especially in the treatment of variability in complex and non-homogeneous data. As noted in [3], [6], current models still present difficulties when segmenting data with mixed features or poorly defined patterns. This situation justifies the need to conduct a systematic review that identifies emerging methods capable of overcoming these challenges, thus promoting a more efficient implementation of data

segmentation in environments characterized by high volumes of information and diversity.

This systematic review identified the main challenges and approaches used in implementing data segmentation using machine learning techniques. The objective was to synthesize the most relevant methodological barriers, analyze the solutions proposed in recent literature, and identify gaps for future research. This article is organized as follows: Section II describes the methodology employed, including the design of the research question, the selection criteria, and the search protocol. Section III presents the results of the review, accompanied by a bibliometric and thematic analysis. Section IV offers a critical discussion of the findings, highlighting the main limitations detected in the field. Finally, Section V presents the main conclusions along with recommendations for future research.

II. METHODOLOGY

A. PICO method and PRISMA protocol

To explore the main challenges in implementing data segmentation using machine learning, the following research question was formulated: What are the main challenges in implementing data segmentation using machine learning methods? This question was designed using the PICO methodology, which allowed the identification of essential components (see Table 1) and their association with relevant keywords to perform a systematic search in the literature. For example, in the Problem component, studies addressing data segmentation in environments with high dimensionality were included, while in the Intervention component, specific methods such as K-means and neural networks were prioritized. Using this methodology, a robust search equation was built (see Table 2), which was executed in the SCOPUS database.

The study addressed the research question using the PRISMA methodology, which rigorously guided the study selection process through inclusion and exclusion criteria, and a detailed review of titles, abstracts, and even full texts. For example, studies published before 2020 or in languages other than English or Spanish were discarded during the process, which allowed the review to be focused but could introduce temporal and linguistic bias. In addition, the manual analysis of 983 initial results, which led to the inclusion of 139 articles and the incorporation of 44 studies after careful interpretation of their abstracts, could have been subject to selection and interpretation biases, which were mitigated by applying

predefined criteria and, where possible, cross-checking between reviewers (see Table 3).

a.1. Research questions

The research question was formulated and structured according to the components of Problem, Intervention, Context and Outcome (PICO) [7], in relation to the challenges associated with the implementation of data classification through machine learning. Table 1 details the PICO structure used to design this research question. On the other hand, Table 2 provides the set of keywords and justifications used in the construction of the search equation.

TABLE I
PICO SUMMERY

Problem	Literature on data segmentation with Machine Learning.
Intervention	Methods applied in data segmentation
Context	Sectors where data segmentation has been implemented.
Result	Results after implementing Machine Learning in data segmentation.

TABLE II
RESEARCH QUESTIONS

RQ	Research Question	Motivation
RQ1	What challenges have been encountered when implementing data segmentation using Machine Learning?	Significant work on segmentation challenges is identified.
RQ2	What Machine Learning methods have been most frequently used to segment data?	Research on applied segmentation techniques has been explored.
RQ3	In which sectors have data segmentation methods using Machine Learning been applied?	Studies in domains with implemented segmentation are analyzed.
RQ4	What were the results of implementing Machine Learning in data segmentation?	Results obtained after applying Machine Learning are examined.
General search equation (total of 983 research articles) (TITLE-ABS-KEY ("Data quality" OR "Implementation challenges" OR "Machine Learning" OR "Data segmentation issues" OR "Data noise") AND TITLE-ABS-KEY ("Clustering algorithms" OR "Neural networks" OR "K-means" OR "Decision trees" OR "DBSCAN") AND TITLE-ABS-KEY ("Retail industry" OR "Healthcare data" OR "Financial sector" OR "Marketing analytics" OR "Telecommunications") AND TITLE-ABS-KEY ("Model performance" OR "Scalability improvements" OR "Accuracy" OR "Computational efficiency" OR "Algorithm effectiveness"))		

To ensure a thorough analysis, not only the titles and abstracts, but also the introductions and the full content of each study were evaluated. This procedure was carried out by strictly applying the established criteria detailed in Table 3.

a.2. Inclusion and exclusion criteria (PRISMA)

Along these lines, 6 inclusion and 5 exclusion criteria were established, which facilitated the selection of 139 relevant documents for the realization of this SLR.

The inclusion criteria were the following:

a) Research and review articles published between 2020 and 2024.

b) Studies focused on the challenges of implementing data segmentation through machine learning.

c) Open access publications.

d) Documents in English or Spanish.

e) Articles related to the areas of engineering, computer science or mathematics.

f) In the case of duplicates, only one of them was considered.

For the exclusion criteria, the following were used:

a) Articles published before 2020.

b) Publications that do not directly address the challenges of data segmentation through machine learning.

c) Documents that do not have open access.

d) Articles in languages other than English or Spanish.

e) In the case of duplicates, additional ones were discarded.

During the initial analysis in SCOPUS, 983 results were identified. The use of Boolean operators confirmed the absence of duplicates, eliminating the need to exclude documents for this reason. After applying the inclusion and exclusion criteria, 139 relevant articles were selected, discarding 844 records that did not meet the established criteria.

In addition, an exhaustive manual review was carried out, incorporating 44 studies after carefully interpreting their abstracts, recognizing that this process could be subject to selection and interpretation biases. To mitigate these biases, predefined criteria were applied and the consistency of the selection between reviewers was verified. These documents were included in the results of the systematic review. Table 3 summarizes the inclusion and exclusion process applied to the studies, whereas Figure 1 graphically represents the selection process following the PRISMA methodology, illustrating the flow of records through the different phases of identification, screening, eligibility, and inclusion.

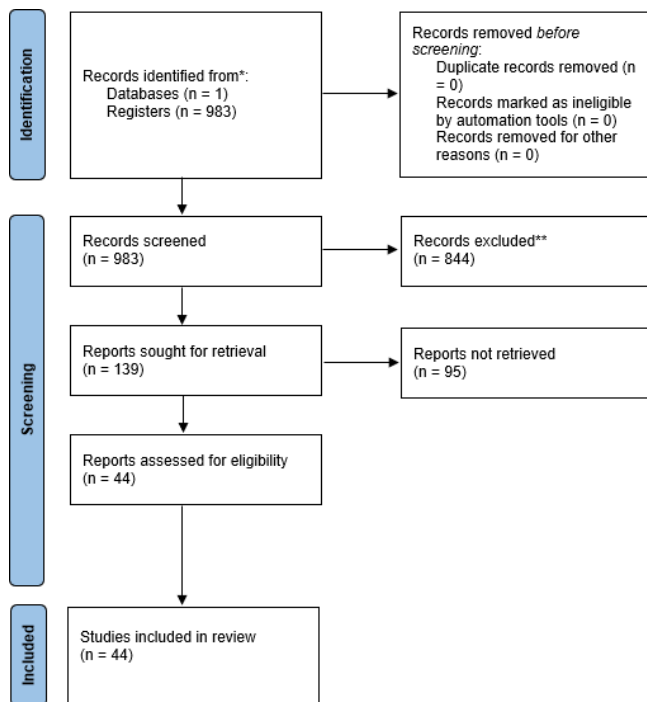


Fig. 1 PRISMA Diagram

III. ANALYSIS OF RESULTS

A. Bibliometric Analysis

Bibliometric analysis based on the SCOPUS database shows that the issue of challenges in implementing data segmentation using Machine Learning has gained relevance since the beginning of the 21st century. Figure 2 illustrates the production of academic literature on this topic over time.

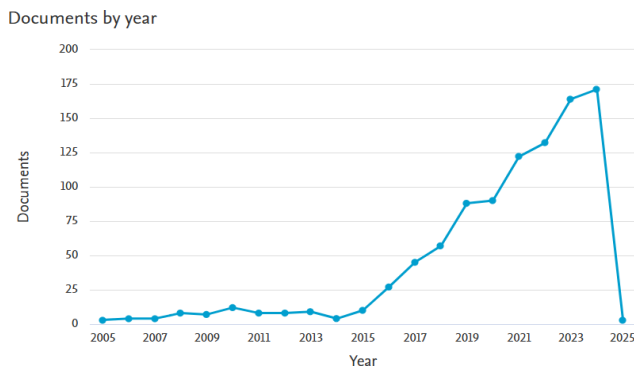


Fig. 2 Production of academic literature.

Likewise, as seen in Figure 3, India emerges as the country with the greatest contribution to this topic, followed by China, the United States, the United Kingdom, Saudi Arabia, Canada, Malaysia, Pakistan, Italy and among others.

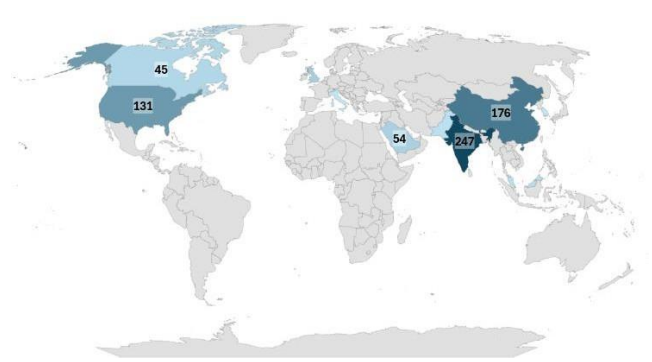


Fig. 3 Countries with the greatest contribution of literature.

Figure 4 presents a network visualization that highlights Machine Learning as the central concept, connecting multiple key topics such as telecommunications, prediction, data analytics, and decision support systems. These concepts reflect the breadth of applications of Machine Learning, spanning areas such as telecommunications traffic, wireless network analysis, logistic regression, and random forests, demonstrating the interrelationship of this technology in sectors such as healthcare, retail, and trend prediction.

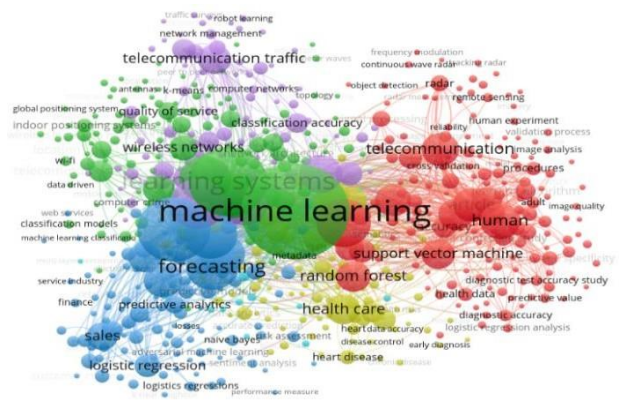


Fig. 4 Network visualization.

Figure 5 shows an overlay visualization that highlights Machine Learning as the central node, around which key themes such as artificial intelligence, prediction, and telecommunications unfold. The color gradation, from blue to yellow, represents the temporal growth and evolution of these concepts between 2018 and 2023.

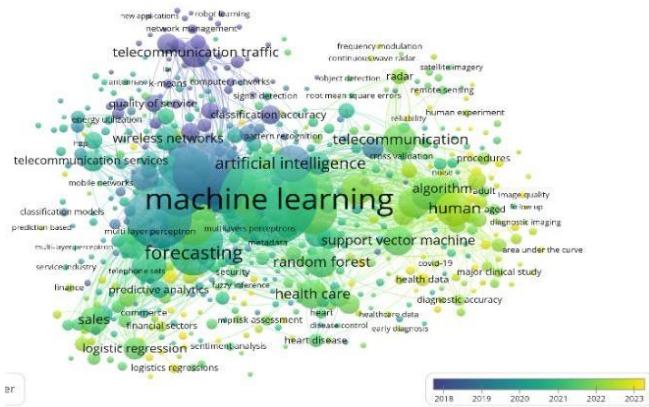


Fig. 5 Overlay visualization.

The analysis of the 44 documents selected in this SLR allowed to answer the questions formulated using the PICO methodology and to define key concepts on data segmentation and machine learning. Data segmentation is understood as the process of grouping heterogeneous sets into meaningful categories to identify relevant patterns using machine learning techniques. This methodology has been used to classify consumer behavior, analyze network traffic data and organize medical data, achieving personalized services, improving clinical diagnoses and optimizing the prediction of customer abandonment (churn) [12], [14], [16], [21], [24], [27], [29], [35], [37], [40].

Machine learning is also a branch of artificial intelligence that uses algorithms to detect complex patterns and develop predictive models from historical data. Its implementation allows to automate the classification and analysis of data in sectors such as health, network traffic and customer behavior. This translates into improvements in diagnostic accuracy, decision optimization and prediction of future events, as evidenced by multiple studies [11], [24], [25], [27], [32], [33], [35], [37], [40], [42].

RQ1: What challenges have been encountered when implementing data segmentation using Machine Learning?

The implementation of data segmentation through machine learning presents a series of significant challenges that hinder its effectiveness. In particular, one of the main problems lies in the high complexity and dimensionality of the data. Handling data sets with these characteristics generates difficulties in parameter adjustment and in the selection of relevant features. This problem has been previously evaluated in sectors such as health and telecommunications, where the volume and diversity of the data increase the complexity of the analysis, directly affecting the accuracy of the models [8], [10], [11], [35], [36], [42].

On the other hand, the variability and noise present in the data constitute another important challenge. Previous studies have highlighted that the high variability of the signals and the

presence of noise in the patterns make accurate segmentation difficult, which reduces the effectiveness of the models. This phenomenon has been widely discussed in the literature, especially in IoT and image analysis contexts, where dynamic changes in signals increase false positive rates and decrease the overall accuracy of algorithms [9], [13], [17], [18], [19], [27], [43], [44].

Similarly, studies reveal that imbalanced data and inconsistent quality issues represent significant obstacles. The presence of imbalanced data, along with incomplete, duplicated, or scattered records, affects the accuracy of model training. In applications such as churn prediction and medical records analysis, inconsistent data quality limits the ability of algorithms to generate effective segmentations [24], [25], [28], [31], [32], [33], [34], [41].

Similarly, domain-specific challenges vary depending on the application context, such as intrusion detection, disease diagnosis, consumer behavior analysis, and telecommunication coverage predictions. These problems have not been thoroughly studied due to data variability, heterogeneity, and pattern complexity, which affects the performance of models in sectors such as e-commerce, healthcare, and telecommunication [12], [14], [15], [16], [20], [21], [22], [23], [26], [29], [38].

Furthermore, scalability and integration issues emerge when handling large volumes of data and integrating heterogeneous data sources. While several theories have been proposed to address the need to scale models and switch between different segmentation techniques, a number of unresolved issues remain related to maintaining performance and processing efficiency in complex environments. This is especially relevant in systems that require integrating data from various devices and information sources [30], [37], [39], [40].

According to Figure 6, the challenges are distributed in: domains (30%), imbalance-quality (22%), variability-noise (21%), complexity-dimensionality (16%) plus scalability-integration (11%). Finally, this evidence suggests exploring methodologies capable of facing these difficulties.

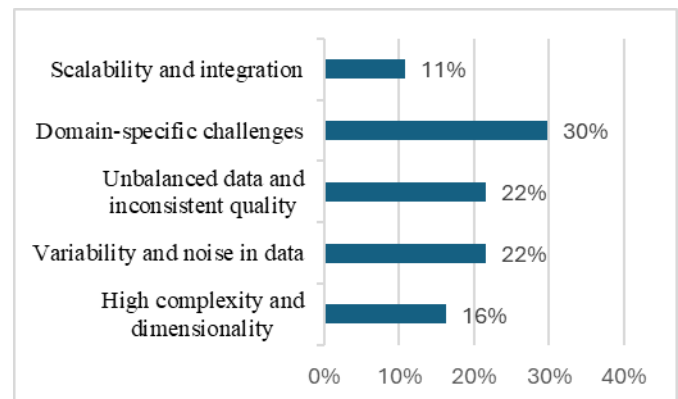


Fig. 6. Challenges in implementing segmentation using M. L.

RQ2: What Machine Learning methods have been most frequently used to segment data?

The literature review shows that in data segmentation analysis using machine learning, deep learning-based models have demonstrated outstanding performance in multiple contexts. Previous research has employed techniques such as convolutional neural networks (CNN), LSTM, stacked autoencoders, and other deep learning models. These methods have been used in various studies to evaluate data segmentation in areas such as traffic analysis, churn prediction, and anomaly detection in medical images. Specific applications such as deep transfer learning with ResNet-50, used for segmentation of areas susceptible to subsidence, and denoising autoencoders, used in segmentation of sequential data, are also part of this group [8], [15], [17], [27], [29], [33], [34], [37], [42].

In parallel, Ensemble learning methods such as Random Forest, Gradient Boosting, XGBoost and LightGBM have been implemented in several studies to improve data segmentation. These methods are frequently combined with optimization techniques such as Bayesian tuning and sampling techniques such as SMOTE. The literature indicates that these combinations have been successfully applied in the classification of intrusions in Internet of Things (IoT) environments, as well as in the segmentation of clients and cellular network data. Furthermore, the use of advanced techniques such as Data Augmentation has allowed to increase the predictive accuracy in large volumes of data [9], [11], [16], [20], [24], [25], [35], [44].

Similarly, in the field of traditional classification methods, algorithms such as Support Vector Machine (SVM), logistic regression, decision trees, Naïve Bayes and KNN stand out in numerous studies. Several authors have recognized the effectiveness of these methods to segment data in various applications. In previous projects, these methods have been successfully applied to segment data in areas such as medical diagnosis, intrusion detection, and customer segmentation in the financial and telecommunication sectors. Furthermore, combinations of these methods, such as decision trees and SVM, have been used to optimize classification and obtain better results in different scenarios [10], [12], [14], [18], [21], [22], [28], [31], [38], [40], [43].

Additionally, several studies have implemented techniques focused on optimization and dimensionality reduction. Algorithms such as PCA, Ridge regression, and Damped Least-Squares have been employed to reduce noise and improve accuracy in predictive models. The literature suggests that the implementation of RFE for feature removal, together with optimization methods such as Levenberg- Marquardt and Bayesian regularization, has been crucial to improve segmentation in complex data analysis [19], [23], [30], [36], [39].

In this regard, hybrid methods and combination of methods have been explored to address complex data segmentation problems. Previous research has implemented fused models such as ACBL-TCN, Deep transfer learning, and

spatiotemporal techniques such as HSTNet in the analysis of RSS signals, interferograms, and other complex data. These methods combine features of different techniques to maximize segmentation accuracy and effectiveness in specific applications, achieving robust results in contexts that require advanced and adaptive models [13], [26], [32], [41].

Figure 7 presents 30% for traditional methods, 24% Deep learning, 22% Ensemble, 13% optimization-reduction and 11% hybrid. Thus, a significant methodological variety emerges.

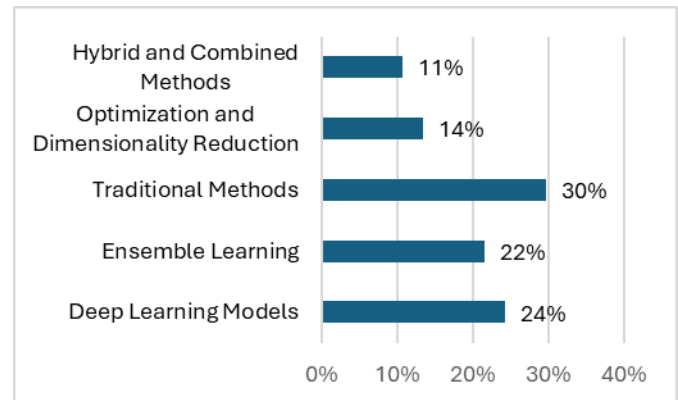


Fig. 7. Machine Learning methods for segmenting data

RQ3: In which sectors have data segmentation methods using Machine Learning been applied?

Previous studies have shown that machine learning data segmentation methods are applied in a variety of sectors, addressing specific problems and contributing to technological advancement in each of them. In the telecommunications sector, there have been numerous studies focused on predictive analysis of customer churn, 5G network coverage optimization, and user retention. For example, machine learning techniques are used to analyze subscriber profiles, predict user behavior, and improve service quality. These methods have been applied in areas such as mobile networks, multimedia streaming services, and retention strategies in highly competitive markets. Specific applications include cellular traffic prediction and network resource allocation, focused on improving user experience and model accuracy [16], [18], [20], [21], [23], [24], [25], [27], [28], [33], [34], [35], [40], [44].

In the healthcare and medical diagnostics field, the literature related to remote patient monitoring and prediction of health conditions using data from wearable devices has grown significantly. Previous research suggests that these methods facilitate early detection and monitoring of chronic diseases such as diabetes, cardiovascular disease, and cancer. Furthermore, advanced imaging techniques and intelligent systems have been employed to improve diagnostic accuracy and continuous patient monitoring, allowing for more personalized and effective care [11], [14], [15], [29], [30], [31], [32], [36], [37], [38], [39], [41].

Regarding network security and cybersecurity, studies of intrusion detection and data protection in environments such as IoT, WSN, and UAV are well documented. Several authors have recognized that these methods are essential to secure communication networks, improve real-time threat detection, and strengthen data security in connected systems. This approach is particularly relevant in sectors that require high reliability and protection against cyberattacks, with an emphasis on network security monitoring systems and traffic analysis [9], [10], [13], [22], [36], [42].

On the other hand, in consumer behavior analysis and e-commerce, several recent studies have indicated that changes in purchasing patterns, especially during the COVID-19 pandemic, have been significant. Researchers have made important contributions by analyzing customer satisfaction and identifying consumption patterns. These investigations focus on developing marketing strategies to improve user experience and adjust business strategies to new trends [12].

In the field of geosciences, agriculture and remote sensing, an extensive literature has been developed on crop classification, ground deformation monitoring and change detection in temporal images. This approach includes applications that use satellite data to improve accuracy in environmental studies and optimize agricultural practices [17], [19], [26].

In the financial and banking field, significant contributions have been made in the application of data segmentation methods in financial services. Authors address problems such as cost prediction, leasing decisions, and optimization of customer retention strategies in the banking sector. These investigations focus on the use of predictive models to improve decision making and adjust business strategies in highly competitive markets [8], [28].

Figure 8 shows telecommunications (37%), healthcare (31%), cybersecurity (16%), geosciences-agriculture (8%), finance-banking (5%) plus e-commerce (3%).

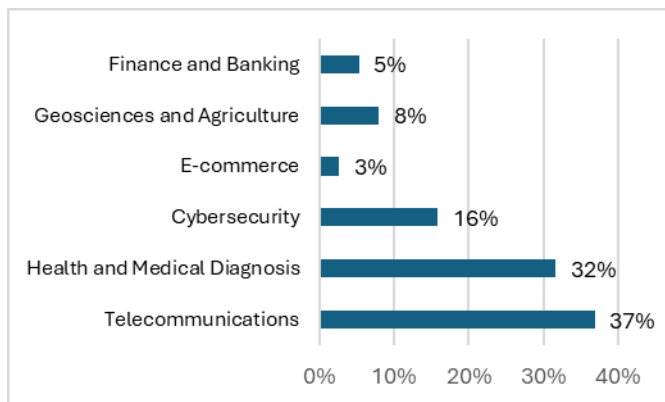


Fig. 8. Sectors for implementing data segmentation using M.L.

RQ4: What were the results of implementing Machine Learning in data segmentation?

First, several studies have shown a significant focus on improving predictive accuracy and efficiency through advanced machine learning techniques. Several authors have recognized that methods such as Deep learning, Ensemble learning and hyperparameter optimization have allowed substantial improvements in key metrics such as precision, AUC-ROC and recall. These improvements have been applied in sectors such as telecommunications, intrusion detection and crop classification, where error rates have been reduced and model robustness increased [8], [10], [11], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [26], [27], [28], [33], [36], [40], [42].

In the field of health and medical diagnosis, previous research has shown that the use of machine learning has optimized patient monitoring and improved diagnostic accuracy in chronic diseases. For example, studies have shown progress in reducing errors and improving the efficiency of health data analysis, using models that predict conditions such as arrhythmias, stress levels, and heart rate. This approach has facilitated early diagnosis and enabled more effective personalization of medical treatment [11], [14], [15], [29], [30], [31], [32], [37], [38], [39], [41].

Similarly, in the telecommunications sector, significant contributions have been made in churn prediction and customer analysis, using techniques such as Ensemble learning and hyperparameter tuning to improve accuracy in identifying customers at risk of churn. These methods have enabled the development of proactive strategies that optimize customer retention, improving business decisions and personalization of services in highly competitive markets [16], [20], [24], [28], [33], [34], [35], [44].

Regarding intrusion detection and network security, previous studies have highlighted that robust models have been implemented that achieve accuracy rates above 90%. Advanced techniques such as DLSA and ResNet-50 have proven effective in reducing false positives and increasing real-time detection, improving security in IoT systems, UAVs, and other connected environments [9], [10], [22], [30], [36], [42].

Likewise, studies related to monitoring systems and quality of experience (QoE) in telecommunications have shown advances in prediction accuracy. It has been reported that the use of data augmentation techniques has optimized the performance and coverage of cellular networks, thus improving user experience in streaming services [23], [25], [27], [35], [40].

Furthermore, some authors have suggested that the use of hybrid techniques and Ensemble methods has improved accuracy in complex classification tasks. These studies demonstrate their effectiveness in areas such as the detection of anomalous traffic, the identification of drowsiness in drivers and the personalization of medical treatments, managing to increase the robustness of the models and reduce

the costs associated with incorrect diagnoses [26], [28], [34], [37], [41], [43].

Figure 9 presents accuracy-efficiency (38%), health (19%), retention (14%), intrusions-security (10%), complex classification (10%) plus networks-QoE (9%).

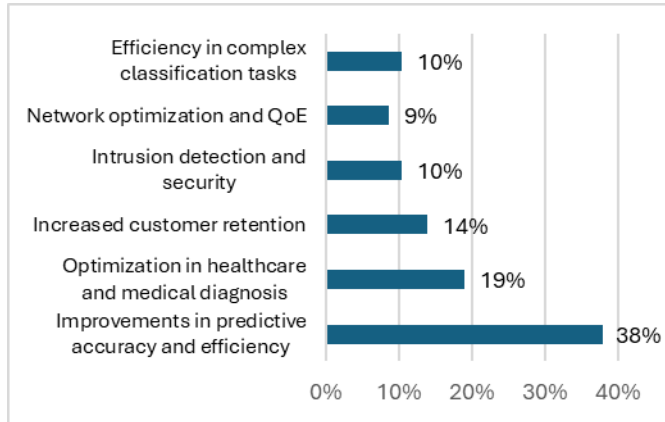


Fig. 9. Results of implementing data segmentation using ML.

IV. DISCUSSION OF RESULTS

Current SLR has demonstrated considerable achievements in implementing data segmentation using machine learning. However, it has also revealed important gaps and unresolved issues that limit its effectiveness in various contexts. Among the most notable findings, the challenges associated with high dimensionality, noise present in the data and inconsistent quality of the records stand out. Although advanced techniques have been developed to address these difficulties, such as dimensionality reduction using PCA and the use of optimization methods such as SMOTE [19], [23], [30], [36], [39], there are still no standardized solutions that can be applied in a generalized way in all sectors. This lack creates a significant gap in the ability to extrapolate the results obtained in specific contexts, such as health or telecommunications, to other equally complex domains [8], [14], [18], [21], [36].

In this sense, the lack of consensus on the appropriate use of hybrid methods and advanced models remains a considerable limitation. While these methods have proven effective in scenarios such as medical image analysis and intrusion detection [13], [26], [32], [41], their implementation remains ad hoc and lacks a clear methodological framework to guide their design and optimization. This situation is particularly evident in sectors that demand adaptability and precision, such as agriculture and cybersecurity, where the heterogeneous characteristics of the data increase the difficulty of applying predefined models [10], [12], [22]. Likewise, there is a lack of longitudinal analysis that evaluates the behavior of the models in dynamic environments over time, which limits the understanding of their sustainability and adaptive capacity [29], [31], [33].

It should be noted that, in the application sectors, the review shows that advances in health and telecommunications

have been substantial. However, the methods used in these sectors are not always transferable to other contexts. For example, techniques developed to optimize customer retention in telecommunications present difficulties when adapted to environmental monitoring or financial analysis, due to differences in the nature of the data and the objectives of the models [16], [20], [24], [28]. This lack of transferability highlights a critical gap in the literature, as it prevents the full use of innovations in machine learning to address problems in less explored sectors [11], [25], [34].

In addition, the reviewed studies suffer from the absence of standardized metrics that allow for consistent comparison of model performance in different applications [35], [42]. The diversity of metrics used complicates the evaluation of results and the identification of the most effective techniques according to the context. Similarly, much of the literature focuses on specific results without exploring how models perform in highly variable and complex scenarios over time, which reflects another important gap in the available knowledge [17], [26], [38].

Consequently, these gaps limit the ability of practitioners to implement robust and scalable solutions in real-world environments. The results obtained highlight the importance of adapting techniques to the specific characteristics of each domain, but challenges persist in integrating heterogeneous data and scaling models to process large volumes of information [9], [30], [37]. Although emerging technologies, such as cloud computing and advanced hardware resources, offer opportunities to overcome these barriers, their widespread adoption requires additional research to validate their effectiveness in diverse contexts and with infrastructure limitations [14], [27], [40]. In this context, the identified gaps underscore the need for more integrative methods that simultaneously address technical and contextual challenges. Future research should focus on developing methodological frameworks that facilitate knowledge transfer across sectors and on standardizing evaluation metrics that allow for objective comparison of results. Although the findings obtained largely answer the questions raised, unsolved problems and unexplored areas show the urgency of continuing to delve deeper into this field, both from theory and practice, to move towards a more precise, adaptive and universally applicable data segmentation [8], [18], [43], [44].

V. CONCLUSIONS

This SLR has shown that, despite the advances achieved in data segmentation using Machine Learning techniques, critical challenges persist that limit their effectiveness in various sectors. These include high dimensionality, intrinsic data complexity, the presence of noise, variability in signals, inconsistent quality problems, as well as imbalance in records, along with difficulties in scalability and integration of heterogeneous sources. These obstacles impact the accuracy of models, especially in contexts characterized by non-homogeneous data or poorly defined patterns.

The application of Deep Learning algorithms, Ensemble Learning, optimization techniques or other hybrid methods has allowed for substantial improvements in areas such as telecommunications, health, cybersecurity or consumer behavior analysis. The absence of a unified methodological framework that enables the widespread adoption of these approaches, coupled with the lack of standardized metrics to compare results, limits the transferability of these methods to domains with specific characteristics.

It is recommended to promote the development of integrative methodological frameworks that simultaneously address both technical and contextual challenges. It is essential to standardize evaluation metrics in order to facilitate comparisons between studies, promoting the exchange of knowledge between sectors. Future research should explore hybrid approaches that combine advanced techniques with adaptive optimization strategies, incorporating longitudinal analysis to evaluate the performance of models over time in dynamic environments.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the Technological University of Peru for its continued support and for the resources provided during this research. Their dedication to promoting academic excellence has been fundamental to the development of this work.

REFERENCES

- [1] F. Moosavi, H. Shiri, J. Wodecki, A. Wyłomańska, and R. Zimroz, "Article Application of Machine Learning Tools for Long-Term Diagnostic Feature Data Segmentation," *Applied Sciences (Switzerland)*, vol. 12, no. 13, 2022, doi: 10.3390/app12136766.
- [2] I. S. Lebedev, "Adaptive application of machine learning models on separate segments of a data sample in regression and classification problems | Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации," *Informatsionno-Upravliaiushchie Sistemy*, no. 3, pp. 20–30, 2022, doi: 10.31799/1684-8853-2022-3-20-30.
- [3] K. S. Mwitondi, I. Munyakazi, and B. N. Gatsheni, "A robust machine learning approach to SDG data segmentation," *J Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00373-y.
- [4] P. Devaghi and S. Sudha, "Exploratory Data Analysis and Data Segmentation using K means Clustering," in *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2023*, 2023, pp. 145–147. doi: 10.1109/ICACITE57410.2023.10183143.
- [5] W. Mounter, C. Ogwumike, H. Dawood, and N. Dawood, "Machine learning and data segmentation for building energy use prediction—a comparative study," *Energies (Basel)*, vol. 14, no. 18, 2021, doi: 10.3390/en14185947.
- [6] Kitchenham Barbara and Charters Stuart, "Guidelines for performing Systematic Literature Reviews in Software Engineering," 2007.
- [7] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Mar. 2021, doi: 10.1136/bmj.N71.
- [8] T. Zema, A. Kozina, A. Sulich, I. Römer, and M. Schieck, "Deep learning and forecasting in practice: an alternative costs case," in *Procedia Computer Science*, 2022, pp. 2952–2961. doi: 10.1016/j.procs.2022.09.354.
- [9] G.-P. Fernando, A.-A. H. Brayan, A. M. Florina, C.-B. Liliana, A.-M. Hector-Gabriel, and T.-S. Reinell, "Enhancing Intrusion Detection in IoT Communications Through ML Model Generalization With a New Dataset (IDSAL)," *IEEE Access*, vol. 11, pp. 70542–70559, 2023, doi: 10.1109/ACCESS.2023.3292267.
- [10] Y. Li, J. Zhang, Y. Yan, Y. Lei, and C. Yin, "Enhancing Network Intrusion Detection Through the Application of the Dung Beetle Optimized Fusion Model," *IEEE Access*, vol. 12, pp. 9483–9496, 2024, doi: 10.1109/ACCESS.2024.3353488.
- [11] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/9930985.
- [12] F. Safara, "A Computational Model to Predict Consumer Behaviour During COVID-19 Pandemic," *Comput Econ*, vol. 59, no. 4, pp. 1525–1538, 2022, doi: 10.1007/s10614-020-10069-3.
- [13] S. B. Altaf Khattak, Fawad, M. M. Nasralla, M. A. Esmail, H. Mostafa, and M. Jia, "WLAN RSS-Based Fingerprinting for Indoor Localization: A Machine Learning Inspired Bag-of-Features Approach," *Sensors*, vol. 22, no. 14, 2022, doi: 10.3390/s22145236.
- [14] J.-W. Oh, "A Study on Digital Healthcare Service in Big Data Environment: Focusing on Diagnosis of Hyperlipidemia Based on Diagnostic Testing," *Journal of System and Management Sciences*, vol. 12, no. 3, pp. 345–360, 2022, doi: 10.33168/JSMS.2022.0317.
- [15] M. Abu-Alhaija and N. M. Turab, "Automated learning of ecg streaming data through machine learning internet of things," *Intelligent Automation and Soft Computing*, vol. 32, no. 1, pp. 45–53, 2022, doi: 10.32604/IASC.2022.021426.
- [16] L. Saha, H. K. Tripathy, F. Masmoudi, and T. Gaber, "A Machine Learning Model for Personalized Tariff Plan based on Customer's Behavior in the Telecom Industry," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, pp. 171–184, 2022, doi: 10.14569/IJACSA.2022.0131023.
- [17] A. Franczyk, J. Bała, and M. Dwornik, "Monitoring Subsidence Area with the Use of Satellite Radar Images and Deep Transfer Learning," *Sensors*, vol. 22, no. 20, 2022, doi: 10.3390/s22207931.
- [18] C. Huang et al., "Machine Learning-Enabled LOS/NLOS Identification for MIMO Systems in Dynamic Environments," *IEEE Trans Wirel Commun*, vol. 19, no. 6, pp. 3643–3657, 2020, doi: 10.1109/TWC.2020.2967726.
- [19] D. K. Kılıç and P. Nielsen, "Comparative Analyses of Unsupervised PCA K-Means Change Detection Algorithm from the Viewpoint of Follow-Up Plan," *Sensors*, vol. 22, no. 23, 2022, doi: 10.3390/s22239172.
- [20] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," *Algorithms*, vol. 17, no. 6, 2024, doi: 10.3390/a17060231.
- [21] M. Gollapalli et al., "Machine Learning Approach to Users' Age Prediction: A Telecom Company Case Study in Saudi Arabia," *Mathematical Modelling of Engineering Problems*, vol. 10, no. 5, pp. 1619–1629, 2023, doi: 10.18280/mmep.100512.
- [22] R. Shrestha, A. Omidkar, S. A. Roudi, R. Abbas, and S. Kim, "Machine-learning-enabled intrusion detection system for cellular connected uav networks," *Electronics (Switzerland)*, vol. 10, no. 13, 2021, doi: 10.3390/electronics10131549.
- [23] M. Sousa, A. Alves, P. Vieira, M. P. Queluz, and A. Rodrigues, "Analysis and Optimization of 5G Coverage Predictions Using a Beamforming Antenna Model and Real Drive Test Measurements," *IEEE Access*, vol. 9, pp. 101787–101808, 2021, doi: 10.1109/ACCESS.2021.3097633.
- [24] M. Imani and H. R. Arabnia, "Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis," *Technologies (Basel)*, vol. 11, no. 6, 2023, doi: 10.3390/technologies11060167.
- [25] M. K. Banjanin, M. Stojčić, D. Danilović, Z. Čurguz, M. Vasiljević, and G. Puzić, "Classification and Prediction of Sustainable Quality of Experience of Telecommunication Service Users Using Machine Learning Models," *Sustainability (Switzerland)*, vol. 14, no. 24, 2022, doi: 10.3390/su142417053.
- [26] M. Choukri, A. Laamrani, and A. Chehbouni, "Use of Optical and Radar Imagery for Crop Type Classification in Africa: A Review," *Sensors*, vol. 24, no. 11, 2024, doi: 10.3390/s24113618.

- [27] D. Zhang, L. Liu, C. Xie, B. Yang, and Q. Liu, "Citywide cellular traffic prediction based on a hybrid spatiotemporal network," *Algorithms*, vol. 13, no. 1, 2020, doi: 10.3390/a13010020.
- [28] Y. Elyusufi and M. A. Kbir, "Churn Prediction Analysis by Combining Machine Learning Algorithms and Best Features Exploration," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, pp. 615–622, 2022, doi: 10.14569/IJACSA.2022.0130773.
- [29] L. V. Coutts, D. Plans, A. W. Brown, and J. Collomosse, "Deep learning with wearable based heart rate variability for prediction of mental and general health," *J Biomed Inform*, vol. 112, 2020, doi: 10.1016/j.jbi.2020.103610.
- [30] A. An, M. Al-Fawa'reh, and J. J. Kang, "Enhanced Heart Rate Prediction Model Using Damped Least-Squares Algorithm," *Sensors*, vol. 22, no. 24, 2022, doi: 10.3390/s22249679.
- [31] T. H. H. Aldhyani, A. S. Alshebami, and M. Y. Alzahrani, "Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms," *J Healthc Eng*, vol. 2020, 2020, doi: 10.1155/2020/4984967.
- [32] J.-C. Kim and K. Chung, "Multi-Modal Stacked Denoising Autoencoder for Handling Missing Data in Healthcare Big Data," *IEEE Access*, vol. 8, pp. 104933–104943, 2020, doi: 10.1109/ACCESS.2020.2997255.
- [33] L. Saha, H. K. Tripathy, T. Gaber, H. El-Gohary, and E.-S. M. El-kenawy, "Deep Churn Prediction Method for Telecommunication Industry," *Sustainability (Switzerland)*, vol. 15, no. 5, 2023, doi: 10.3390/su15054543.
- [34] N. Almufadi and A. M. Qamar, "Deep Convolutional Neural Network Based Churn Prediction for Telecommunication Industry," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 1255–1270, 2022, doi: 10.32604/csse.2022.025029.
- [35] J. Isabona, A. L. Imoize, and Y. Kim, "Machine Learning-Based Boosted Regression Ensemble Combined with Hyperparameter Tuning for Optimal Adaptive Learning," *Sensors*, vol. 22, no. 10, 2022, doi: 10.3390/s22103776.
- [36] G. Lazrek, K. Chetoui, Y. Balboul, S. Mazer, and M. El bekkali, "An RFE/Ridge-ML/DL based anomaly intrusion detection approach for securing IoMT system," *Results in Engineering*, vol. 23, 2024, doi: 10.1016/j.rineng.2024.102659.
- [37] W. T. Al-Sit, N. A. Al-Dmour, T. M. Ghazal, and G. F. Issa, "IoMT- Based Healthcare Framework for Ambient Assisted Living Using a Convolutional Neural Network," *Computers, Materials and Continua*, vol. 74, no. 3, pp. 6867–6878, 2023, doi: 10.32604/cmc.2023.034952.
- [38] V. P. Muse, D. Placido, A. D. Haue, and S. Brunak, "Seasonally adjusted laboratory reference intervals to improve the performance of machine learning models for classification of cardiovascular diseases," *BMC Med Inform Decis Mak*, vol. 24, no. 1, 2024, doi: 10.1186/s12911-024-02467-6.
- [39] Q. An, P. Szewczyk, M. N. Johnstone, and J. Jin Kang, "Enhancement of Healthcare Data Performance Metrics using Neural Network Machine Learning Algorithms," in *2021 31st International Telecommunication Networks and Applications Conference, ITNAC 2021*, 2021, pp. 172–179, doi: 10.1109/ITNAC53136.2021.9652158.
- [40] C. C. Gonzalez, E. F. Pupo, D. P. Ruisanchez, D. Plets, and M. Murroni, "Three-stages concatenated Machine Learning model for SFN prediction," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, 2021, doi: 10.1109/BMSB53066.2021.9547146.
- [41] A. Sheik Abdullah, V. Naga Pranava Shashank, and D. Altrin Lloyd Hudson, "Disseminating the Risk Factors With Enhancement in Precision Medicine Using Comparative Machine Learning Models for Healthcare Data," *IEEE Access*, vol. 12, pp. 72794–72812, 2024, doi: 10.1109/ACCESS.2024.3400023.
- [42] D. Li, X. Dong, J. Gao, and K. Hu, "Abnormal Traffic Detection Based on Attention and Big Step Convolution," *IEEE Access*, vol. 11, pp. 64957–64967, 2023, doi: 10.1109/ACCESS.2023.3289200.
- [43] H. U. R. Siddiqui et al., "Ultra-Wide Band Radar Empowered Driver Drowsiness Detection with Convolutional Spatial Feature Engineering and Artificial Intelligence," *Sensors*, vol. 24, no. 12, 2024, doi: 10.3390/s24123754.
- [44] M. Afzal, S. Rahman, D. Singh, and A. Imran, "Cross-Sector Application of Machine Learning in Telecommunications: Enhancing Customer Retention Through Comparative Analysis of Ensemble Methods," *IEEE Access*, vol. 12, pp. 115256–115267, 2024, doi: 10.1109/ACCESS.2024.3445281.